

# **Scaling** WGS Association Studies in the **Cloud** with the **Seven Bridges** platform

**David Roberson**  
Community Engagement  
Seven Bridges

**Summer Institute in Statistical Genetics 2019**

<http://bit.ly/datastage-sb>

# Part I - Agenda

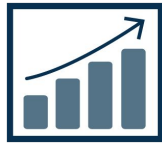
- What is a Seven Bridges Analysis Platform?
- Login to “DataSTAGE powered by Seven Bridges”
- Navigate the Project UI
- Start RStudio Data Cruncher
- In Part II (Friday) we will learn about advanced features for scaling NGS workflows

# Seven Bridges helps researchers do more

- A stable, secure, and highly customizable cloud storage and computing platform
- A user-friendly portal for collaborative analysis of petabytes of public data alongside private data
- An optimized venue for reproducible data analysis using validated tools and pipelines



Easy data  
management



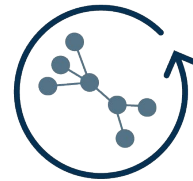
Scalable  
computation



Optimized  
bioinformatics  
algorithms



Secure  
collaboration



Flexible & fully  
reproducible  
methods



Extensible and  
developer-friendly  
platform

# Seven Bridges Public Programs

Infrastructure



**DataSTAGE**  
Powered by Seven Bridges

Partnerships



Genomics  
england



**BloodPAC**  
BLOOD PROFILING • ATLAS IN CANCER



**PCAWG**  
PanCancer Analysis  
OF WHOLE GENOMES

**CH**  
The Children's Hospital  
of Philadelphia®



#### Log in via an external account

Log in via either of the following external accounts:

**eRA Commons\***  
**NIH CIT**

\* To access dbGaP Controlled Data via the DataSTAGE powered by Seven Bridges, log in using the eRA Commons account that is associated with your approved dbGaP data access request. [Learn more.](#)

Log in

Having trouble logging in via eRA Commons?  
Try resetting your password.

#### Log in with your platform account

Use this option if you don't have an external account or don't want to use it.

Username or email

Password

[forgot?](#)

Log in

Chrome is the only recommended browser (July 19)

#### Warning Notice

This is a U.S. Government information system, which may be accessed and used only for authorized Government business by authorized personnel. Unauthorized access or use of this system may subject violators to criminal, civil, and/or administrative action.

All information on this computer system may be intercepted, recorded, read, copied, and disclosed by and to authorized personnel for official purposes, including criminal investigations. Such information includes sensitive data encrypted to comply with confidentiality and privacy requirements. Access or use of this computer system by any person, whether authorized or unauthorized, constitutes consent to these terms. There is no right of privacy in this system.

#### Funding

DataSTAGE powered by Seven Bridges has been funded in whole or in part with Federal funds from the National Institutes of Health, Department of Health and Human Services, under Other Transaction Agreement Nos. 1 OT3 OD025463-01 and 3 OT3

Login using your user name and password on the right side only

- On your dashboard click “create a project” (bottom left)
- Add SISG19 and your name to the project name
- Billing group should be SISG19
- Location should be AWS
- Uncheck “Controlled Data”

+ Create a project View all projects

## Create a project ✕

Name

SISG19-DaveRoberson

Project URL:

https://f4c.sbgenomics.com/u/dave/sisg19-daveroberson 

Billing Group

SISG19 ▼

Location 

AWS (us-east-1) ▼


Execution settings:

**Spot Instances** 

On

**Memoization** **BETA** 

Off

This project will contain **CONTROLLED** Data. 

Cancel

Create

## DESCRIPTION

## Welcome to your new project!

Projects are the core building blocks of the DataSTAGE powered by Seven Bridges Platform. Each project corresponds to a distinct scientific investigation, serving as a container for its data, analysis pipelines, and results. Projects are shared only by designated project members.

### Within your project, you can:

- Start [exploring public datasets](#) straight away
- [Install your tools on the platform](#) and create workflows
- [Upload your own private data](#) and analyze it along with public datasets
- [Collaborate securely](#) with other researchers

Please record the details of your project here, such as its aims, experimental context, and any other ideas that you'd like to share with your project members. Remember that details of each pipeline execution you run on the platform are logged on the task page. This notepad is just for your own notes.

You can also [use markdown](#) here to add formatting to your notes.

Good luck with your research! If you get stuck, take a look at the [Knowledge Center](#)

The Seven Bridges Team

[Add description](#)

## MEMBERS

[🔔 Email notifications](#)dave OWNER

Write, Copy, Execute, Admin

Don't work alone.  
The best research happens in teams.

[+ Invite new members](#)

Share your tools, data, and ideas with collaborators

## ANALYSES

[Tasks](#)[Data Cruncher](#)

Your executions will appear here.  
Before you start, [learn more about them.](#)

## Explore genomics data

Understand complex genomics data with interactive analysis tools.



### Genome Browser

Visualize alignments, SNV/Indels, annotation tracks, check coverage and mismatch, assess alignments and variants

0 files

Open



### Data Cruncher

Analyze and explore data using JupyterLab or RStudio

Open



### Variant Browser BETA

Filter and interpret your annotated data

0 files

Open



## Interactive data science and scientific computing across multiple programming languages.

Your analyses will appear here. [Learn more](#)

Create your first analysis

Create new analysis ✕

✓Basic information ✓Compute requirements

Analysis name


SISG Module 17 WGS RStudio


Environment

**JupyterLab**  
Web-based UI for Project Jupyter

**RStudio BETA**  
IDE for R

Skip wizard Previous **Next**

1 

2 

Create new analysis ✕

✓Basic information ✓Compute requirements


Select an instance type with adequate CPU, memory and storage allocation for your analysis. This can be changed between analysis runs, but not while the analysis is running.


Instance type Suspend time ⓘ On

c3.2xlarge (160GB SSD, 8vCPUs, 15GB RA...  Minutes

Price: \$0.44 per hour

Skip wizard Previous **Start the analysis**

3 

4 

INITIALIZING

2/3

## SISG Module 17 WGS RStudio

t...

Session started on July 22, 2019 10:58 by dave

Environment: RStudio | Duration: Less than a minute

[View Sessions](#)

📄 Copy

✕ Stop

**i** We are preparing your instance, which should not take more than a few minutes. You will be notified when it is ready.

Files

Settings

### Analysis files:

No notebooks

### Produced by this analysis

No files

Select analysis file to preview.

✔ Your instance is up and running! Open the editor to start working on your analysis. ✕

**RUNNING** SISG Module 17 WGS RStudio 

Session started on July 22, 2019 10:58 by dave  
Environment: RStudio | Price: \$0.03 | Duration: 5 minutes [View Sessions](#)

 Copy  Stop  Open in editor

 Your instance is up and running. After 300 minutes of inactivity we will stop the instance and save your analysis. [Learn more](#)

Files Settings

**Analysis files:**

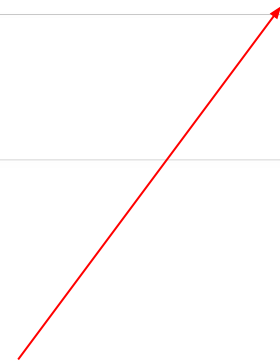
No notebooks

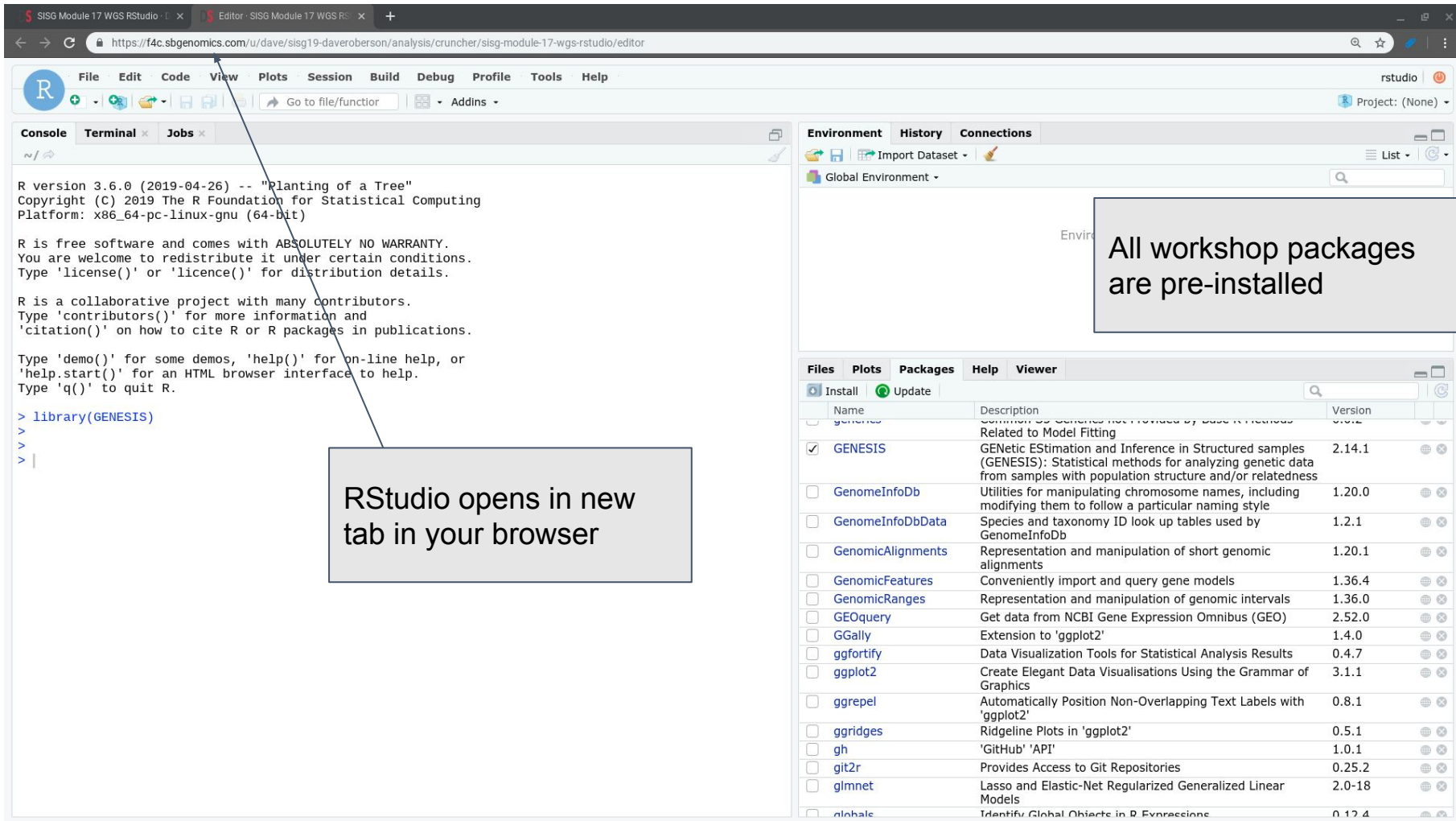
**Produced by this analysis**

No files

Select analysis file to preview.

Click open in editor





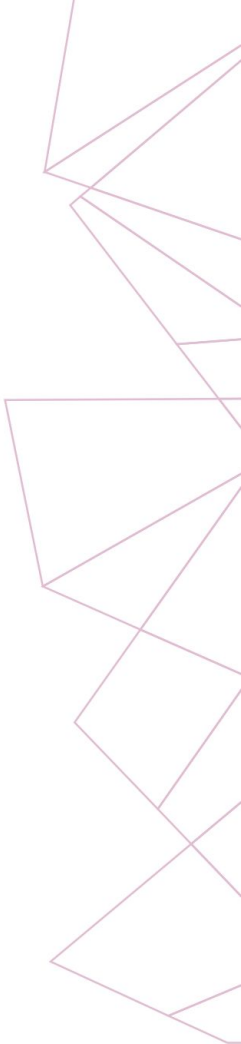
RStudio opens in new tab in your browser

All workshop packages are pre-installed

Name	Description	Version
<input type="checkbox"/> <a href="#">genetics</a>	Common to Genetics not provided by base R functions Related to Model Fitting	0.0.1
<input checked="" type="checkbox"/> <a href="#">GENESIS</a>	GENetic ESTimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness	2.14.1
<input type="checkbox"/> <a href="#">GenomeInfoDb</a>	Utilities for manipulating chromosome names, including modifying them to follow a particular naming style	1.20.0
<input type="checkbox"/> <a href="#">GenomeInfoDbData</a>	Species and taxonomy ID look up tables used by GenomeInfoDb	1.2.1
<input type="checkbox"/> <a href="#">GenomicAlignments</a>	Representation and manipulation of short genomic alignments	1.20.1
<input type="checkbox"/> <a href="#">GenomicFeatures</a>	Conveniently import and query gene models	1.36.4
<input type="checkbox"/> <a href="#">GenomicRanges</a>	Representation and manipulation of genomic intervals	1.36.0
<input type="checkbox"/> <a href="#">GEOquery</a>	Get data from NCBI Gene Expression Omnibus (GEO)	2.52.0
<input type="checkbox"/> <a href="#">GGally</a>	Extension to 'ggplot2'	1.4.0
<input type="checkbox"/> <a href="#">ggfortify</a>	Data Visualization Tools for Statistical Analysis Results	0.4.7
<input type="checkbox"/> <a href="#">ggplot2</a>	Create Elegant Data Visualisations Using the Grammar of Graphics	3.1.1
<input type="checkbox"/> <a href="#">ggrepel</a>	Automatically Position Non-Overlapping Text Labels with 'ggplot2'	0.8.1
<input type="checkbox"/> <a href="#">ggribes</a>	Ridgeline Plots in 'ggplot2'	0.5.1
<input type="checkbox"/> <a href="#">gh</a>	'GitHub' 'API'	1.0.1
<input type="checkbox"/> <a href="#">git2r</a>	Provides Access to Git Repositories	0.25.2
<input type="checkbox"/> <a href="#">glmnet</a>	Lasso and Elastic-Net Regularized Generalized Linear Models	2.0-18
<input type="checkbox"/> <a href="#">globale</a>	Identify Global Objects in R Expressions	0.12.4

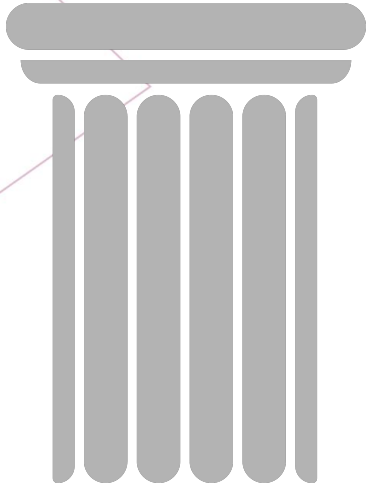
# Part II - Agenda

- Background on DataSTAGE
- Run Single Variant Association at Scale
  - Graphical Data Workflows
    - Clone Public Analysis Pipelines (Hands On)
  - Scaling Association Studies
    - Run association analysis in the cloud (Hands On)
    - Iteration with new parameters (Hand On)
- Time permitting
  - Workflow editor
  - Finding TOPMed Study Data

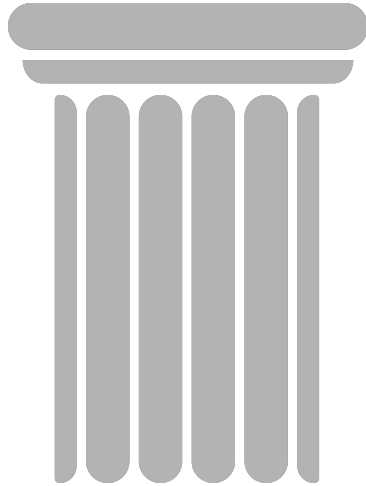


# National Heart, Lung, and Blood Institute (NHLBI) DataSTAGE ecosystem

## Mission



## Vision



The NHLBI DataSTAGE's *mission* is to develop and integrate advanced cyberinfrastructure, leading edge tools, and FAIR data to support the NHLBI research community.

The *vision* for DataSTAGE is to be a community-driven ecosystem implementing data science solutions to democratize data and computational access to advance Heart, Lung, Blood, and Sleep science.



# NHLBI DataSTAGE ecosystem

Several organizations including Seven Bridges are working together to build an ecosystem of services that will host NHLBI datasets in the cloud and allow researchers to access, search, and compute on those data in the cloud.

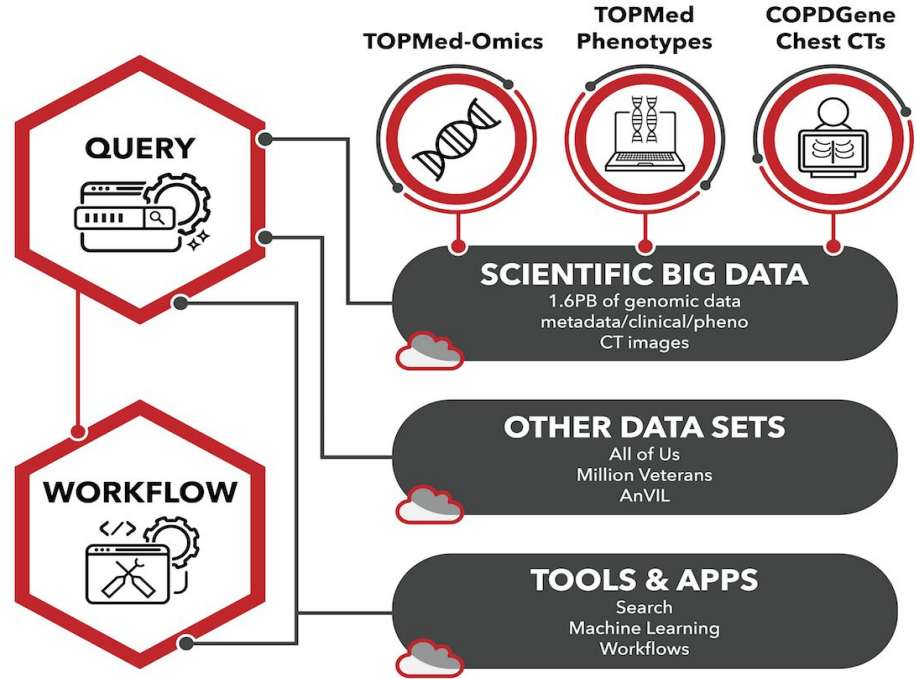
Seven Bridges is building a cloud-based data analysis platform to support accessing the hosted datasets and computing at scale with an emphasis on enabling researchers to perform association studies on WGS data.



# Initially hosting Trans-omics for Precision Medicine (TOPMed) data



AuthZ/AuthN





# DataSTAGE Powered By Seven Bridges Key Features

- Interactive Analysis Data Cruncher
  - R and Jupyter notebooks
- Batch task executor
  - Runs Common Workflow Language (CWL) natively
  - Scatters across and inside of cloud instances
- Graphical tool and workflow editor
- Data Browser
  - Access and use TOPMed Data

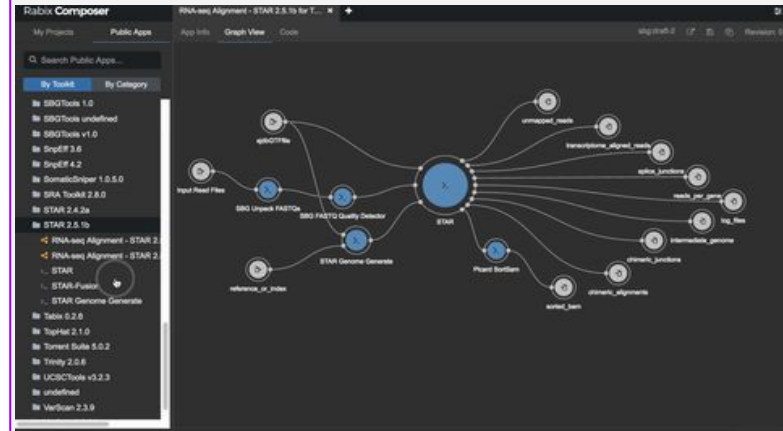
## All Features Include:

- Secure project workspace based collaboration
- Compute where your data is stored

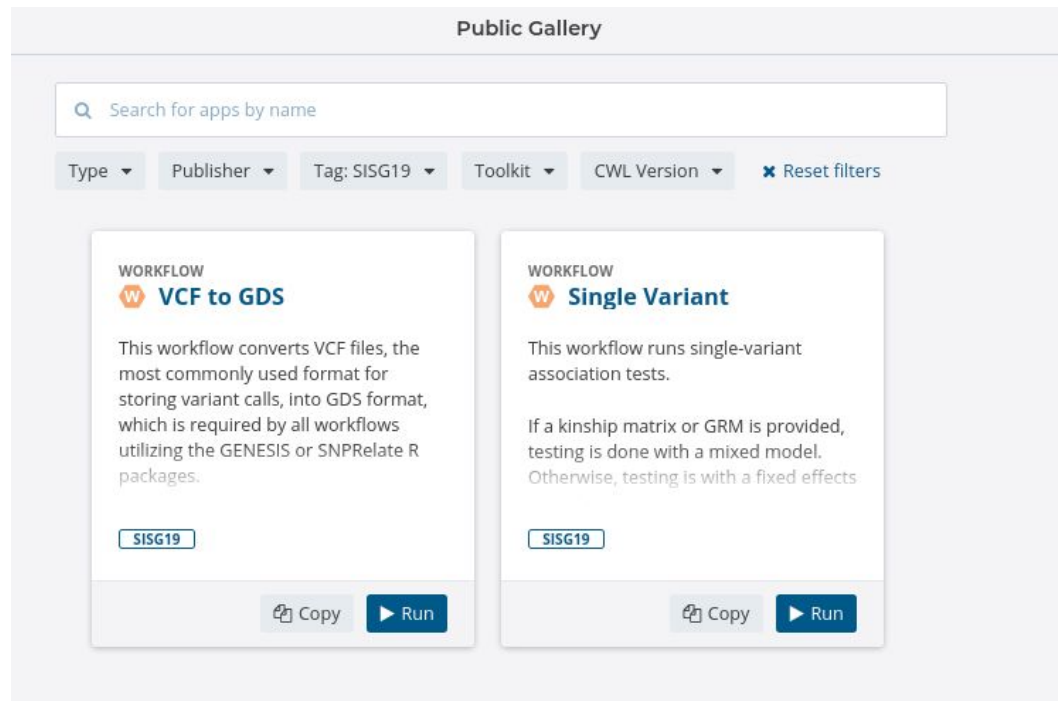
- Portable directed acyclic graph of a scientific workflow
- Easy for anyone to manipulate
- Users can bring their own scripts and tools and make them portable
- Workflows can be nested



## Graphical Representation of Data Workflows



- We will scale 2 pipelines
- Temporarily published in the Public Gallery for this workshop only



The screenshot displays the 'Public Gallery' interface. At the top, there is a search bar with the placeholder text 'Search for apps by name'. Below the search bar are several filter buttons: 'Type', 'Publisher', 'Tag: SISG19', 'Toolkit', 'CWL Version', and a 'Reset filters' button with a close icon. Two workflow cards are visible. The first card is titled 'WORKFLOW VCF to GDS' and includes a description: 'This workflow converts VCF files, the most commonly used format for storing variant calls, into GDS format, which is required by all workflows utilizing the GENESIS or SNPRelate R packages.' It features a 'SISG19' tag and 'Copy' and 'Run' buttons. The second card is titled 'WORKFLOW Single Variant' and includes a description: 'This workflow runs single-variant association tests. If a kinship matrix or GRM is provided, testing is done with a mixed model. Otherwise, testing is with a fixed effects.' It also features a 'SISG19' tag and 'Copy' and 'Run' buttons.

# Hands On Exercise

- Login to the platform and navigate to your working project
- Together we will work on
  - Cloning 2 Public Analysis Pipelines
  - Copying public files to your projects
  - Running VCF 2 GDS workflow
  - Running single variant association workflow